岡山理科大学紀要 第43号 A pp. 47-56(2007)

視野の広さの違いによるAHP強化学習の性能比較

太田真由美・片山　謙吾*・南原　英生*・成久　洋之*

岡山理科大学大学院工学研究科情報工学専攻

*岡山理科大学工学部情報工学科

(2007年10月1日受付、2007年11月2日受理)

1. まえがき

1995年に発生した阪神淡路大震災は，想像を絶する大災害であったため，被災者の救助活動は困難を極めた．このような災害現場において，人間の代わりに迅速に被災者の救助活動を行うことができる，複数の自律ロボットによるマルチエージェントシステム（Multiagent System）[1]の実現が強く求められている[15]．マルチエージェントシステムがおかれる環境として想定されるのは，自律ロボットの活躍が期待される災害現場のように，大規模で複雑（動的・未知）な環境である場合が多い．しかし，そのような環境に適応できるマルチエージェントを設計することは非常に困難である．なぜならば，設計者が予め起こりうる全ての状況を予測し，知識をプログラム化して，エージェントに与えておくことは事実上不可能だからである．よって，各エージェントが自身の経験を通してタスクを達成する方法を学習できる機能を備えていることが望ましいといえる．そのような学習機能として，設計者が目標達成時に与える報酬の設定をするだけで，エージェントが自律的に環境との相互作用を通して，報酬を最大にする適応行動を獲得していく強化学習 (Reinforcement Learning)[5, 14]による機械学習アプローチが注目を集めている[2, 6, 11]．

強化学習は，上述したような動的かつ不確実性を含む環境への対応が期待されているが，環境に対する情報を全く持たず，報酬だけを手がかりに学習を行うため，多くの学習時間を余儀なくされる場合が多い．そのため，現在のところ，実問題への応用や実際的な場面における利用において十分に対応できるとは言い難い．そのような学習の状況を踏まえると，特定の目標を達成するために迅速な対応が要求される応用（e.g., サッカーゲーム）や，安全性や確実性が求められるような応用（e.g., 交通信号制御）では多くの問題を誘発する．これらの現実的な応用においては，試行錯誤を伴う意思決定による行動が行われる限り，迅速な対応が期待できないだけでなく，人命に関わる可能性もある．そのような観点から，強化学習の長所を保ちつつ，現実的により利用しやすい，強化学習をベースとする手法が要求されている．そこで我々は，学習エージェント自身が設定目標を達成するために本来備えておくべき基礎知識を階層化意思決定法（Analytic Hierarchy Process, AHP）で設計し，その基礎知識をAHP器として従来の強化学習エージェントへ導入するAHP強化学習を提案している[6]．

AHP器は，エージェントが認識した環境の状態からAHP器が必要な情報を取り出し，意思決定に利用している．よって，エージェントが認識できる視野の広さにより，AHP器の性能に違いが生じてくることが予想される．そこで本研究では，AHP強化学習の視野の広さの違いによる学習性能の検討を行う．対象問題として，これまでマルチエージェント強化学習の研究で対象とされてきた単一タスクの問題（例：追跡問題[6]）ではなく，現実問題において多く存在する複数タスクの問題[12]を対象とし，複数タスク問題としてレスキュー問題を使用する．

2. 強化学習

2.1 枠組み

強化学習エージェントは，環境の状態を認識し，それに対してエージェントが可能な行動群の中から行動を一つ選択して実行する．この状態認識と行動を繰り返した結果，目標状態に達したとき，環境から報酬が与えられる．エージェントは報酬をもたらす行動を優先するように環境への適応を目指す．

2.2 Profit Sharing

本研究では，強化学習手法として，マルチエージェント環境において有効とされている[1] Profit Sharingを用いる．Profit Sharingは，報酬に至るまでのエピソードにおける状態 s と実際に行った行動 a の対から

[1]複数の自律エージェントが相互に作用しあいながら問題を解決するシステムをマルチエージェントシステムという．また，エージェントとは，行動を行うことによって，自分がおかれている環境に対して影響を与えることのできる自律的主体を指す[13]．

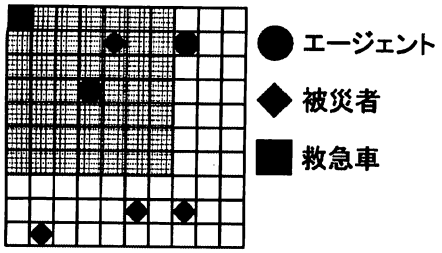

Fig.1 2次元格子状環境

なるルール系列を記憶しておき，報酬が得られたときにそれまでの系列上のルールを一括して強化する学習方法である．ルール系列は次式を用いて強化する．

$$w(s_i,a_i) \leftarrow w(s_i,a_i) + f(r,i) \quad (1)$$

$$f(r,i) = \beta^{W-i} r \quad (2)$$

ここで，$w(s_i,a_i)$ はエピソード系列上の i 番目のルールの重み，f は強化関数，r は報酬値，$\beta(0 \leq \beta \leq 1)$ は報酬割引率，W はエピソードの最大長である．

2.3 ルーレット選択

本研究で使用するエージェントの行動選択法は，Profit Sharing の学習過程において，経験的に良い性能を示すことが知られている[1] ルーレット選択法を使用する．ルーレット選択法は，ある状態 s において，各行動の重み $w(s,a_t)$ を全ての行動の重みの合計 $\sum w(s,a_t)$ で割り，確率 $P(a_t|s)$ を求め，その確率により行動を決定する方法である．

$$P(a_t|s) = w(s,a_t) / \sum w(s,a_t) \quad (3)$$

3. レスキュー問題

レスキュー問題とは，エージェントがある環境中に存在するすべての被災者を救急車に運び込むことを目標とする問題である．この問題には，被災者を探し抱えて，抱えた被災者を救急車に運び込むという連鎖的な複数のタスクが存在する．以下，本研究で利用するレスキュー問題の設定に関して記述する．

Fig.1 に示す $n \times n$ の2次元格子状の環境を設定し，格子の外枠を障壁とする．この環境に救急車 N_{amb} 個を左上端に固定配置，被災者 N_v 個をランダム配置し，エージェント N_A 個すべての初期位置を救急車と同じマスとする．各エージェント $A_j(j=1,\cdots,N_A)$ は同時に行動し，上下左右に1マス進むまたは停止の行動を選択することができる．エージェントは，被災者のマスと同じマスになったとき，被災者を抱えることができる．また，救急車のマスと同じマスになったとき，抱えている被災者を救急車に運びこむことができる．エージェントが行動した単位時間を1ステップとする．

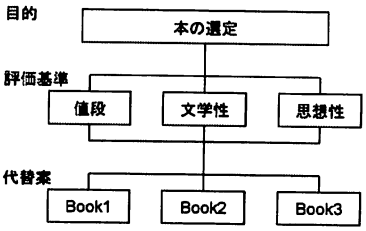

Fig.2 本の選定に関する階層構造の一例

エージェントの視界（Fig.1 網掛け）は $m \times m$ で与え，その視界内に存在する他のエージェント，救急車，被災者，障害物，障壁を認識することができる．レスキュー問題では，被災者を探し抱えたとき，被災者を救急車に運び込んだとき副目標達成となる．そして，すべての被災者を救急車に運び込んだとき主目標達成となる．主目標を達成するまでを1エピソードとする．

4. 階層化意思決定法

ある問題を解決するために，我々人間が行う意思決定は主観的な観点にもとづいてなされる場合が多い．階層化意思決定法（Analytic Hierarchy Process, AHP）は，問題を解決するための代替案がいくつか与えられ，それらの代替案の中から一つを選択する際に主観的な評価に頼らざるを得ない状況において利用される手法である．オペレーションズリサーチの分野などを中心として，さまざまなタイプの AHP やその発展法が精力的に研究されているが，本論文で扱う AHP の構造は典型的なものである．

典型的な AHP の処理の流れについて述べる．AHP は，（1）対象とする意思決定の問題を階層構造に分解する．一般に階層構造は，問題を「目的（goal）」「評価基準（criteria）」「代替案（alternatives）」の関係で捉えることで構築される．問題の「目的」と「代替案」は予め与えられる．「評価基準」は各要素間の一対比較の際に相対的な重み付けを行うための評価の基準であり，より複雑な評価基準を複数の階層を用いることで構築する場合もある．（2）各階層の要素間の一対比較にもとづき一対比較表を作成し重み付けを行う．それらの重み付けにもとづき，（3）階層全体の重み付けを行うことで，各代替案の重要度を算出し，「目的」に対する代替案の優先度を定量的に決定する．

例として，三つの本（Book1，Book2，Book3）の中からどの本を買うべきかを選定する問題について考える．この例の場合，目的はどの本を買うかという「本の選定」となり，代替案は三つの本である「Book1」「Book2」「Book3」が与えられる．評価基準として「値段」や「文学性」，「思想性」を挙げたとすると，その階層構造は Fig.2 のようになる．

Table 1 「値段」に関する主観的判断による一対比較表の一例

値段	Book1	Book2	Book3
Book1	同じ	やや安い	非常に安い
Book2	やや高い	同じ	やや高い
Book3	非常に高い	やや安い	同じ

Table 2 AHP における一対比較値

一対比較値	定義
1	両方の要素が同じぐらい重要
3	行の要素の方が列の要素より少し重要
5	行の要素の方が列の要素よりかなり重要
7	行の要素の方が列の要素より非常に重要
9	行の要素の方が列の要素より極めて重要
上の数値の逆数	重要でない場合に用いる

例えば，人間の主観的判断により「値段」の評価基準で各本を評価した場合に Table 1 に示す結果が得られたとする．ここでは,「安い」という評価が評価値としては高く,「高い」という評価が評価値としては低いとする．しかしながら，このような曖昧な表現では各本の評価を数量化することは困難である．そこで AHP では，曖昧な表現を Table 2 に示すような整数値（一対比較値）に置き換える．Table 1 の「やや安い」の場合は **3**,「非常に安い」の場合は **7**，というように置き換える．これにより，Table 1 は Table 3 のようになる．同様に，すべての評価基準にもとづいてそれぞれの一対比較表を作成する（本例では,「値段」の他に「文学性」「思想性」の一対比較表ができる）．さらに評価基準（「値段」「文学性」「思想性」）の間の一対比較も行う．この場合は 1 つ上の階層の要素（本例では,「本の選定」）に関して一対比較表を作成する．

一対比較表を作成する際，次の 2 点に注意する必要がある．1 つ目は対角要素を 1 にすること，2 つ目は，要素 i からみた要素 j の一対比較値を v_{ij} とする場合，その対角要素の値は $1/v_{ij}$ とすることである．よって，一対比較表の対角線より上の評価を定めることによって対角線より下の評価が可能になり，一対比較表が作成される．

作成された一対比較表から各項目の重みを計算する方法として幾何平均法がよく知られている．幾何平均法は各項目の行を幾何平均し，重みは幾何平均値の合計が 1 になるように正規化することで得られる．つまり，一対比較表の i 行目の幾何平均値 GM_i を $\sqrt[k]{v_{i1} \cdot v_{i2} \cdots v_{ik}}$ により算出する．ここで，$v_{ij}(i, j \in \{1, ..., k\})$ は i 行 j 列にある各要素の比較値である．各幾何平均値の和 $sum = \sum_{i=1}^{k} GM_i$ を求め，i 行目の重み $w_i = GM_i/sum$ をそれぞれ算出する．

Table 3 の場合，各行の幾何平均は Book1:2.76, Book2:0.48, Book3:0.75 となり，幾何平均の和は 3.99 となる．この和で各行の幾何平均値を割ると Book1:0.69, Book2:0.12, Book3:0.19 が求められ，値

Table 3 「値段」に関する一対比較表の一例

値段	Book1	Book2	Book3	幾何平均 GM	重み w
Book1	1	3	7	2.759	0.6908
Book2	1/3	1	1/3	0.481	0.1204
Book3	1/7	3	1	0.754	0.1888

Table 4 目的及び各評価項目に対する一対比較表

(a) 目的に対する一対比較表

目的	重み w
値段	0.1428
文学性	0.4286
思想性	0.4286

(b) 各評価項目に対する一対比較表

	重み w		
	値段	文学性	思想性
Book1	0.6908	0.0976	0.1350
Book2	0.1204	0.3879	0.2808
Book3	0.1888	0.5145	0.5842

段の場合では Book1 が最も好ましいことが示される．上述したように，この操作をすべての評価基準について行い，さらにどの評価項目が重視されるかの判断も同様の操作を行うことで導き，最終的にどの代替案を選定すべきかの判断を下す．

最終的な代替案の重みは，代替案が有する重みと目的に対する各評価項目の重みを掛け合わせ，それらを加算することで求められる．以下では，Book1 を例にとり，問題に対する最終的な重みの算出手順を示す．各項目ごとに判断を行った結果，Table 4 が得られたとする．そこで Table 4 の (a) から目的に対する重みと，(b) から Book1 の各評価項目に対する重みを用いると，$0.1428 \times 0.6908 + 0.4286 \times 0.0976 + 0.4286 \times 0.1350 = 0.1923$ となり，この値が Book1 の最終的な重みとなる．このような処理を Book2，Book3 ともに行うと Table 5 が得られる．よって，Table 5 の結果から Book3 を選択することが適切であると判断を下させる．このように AHP では，主観的な判断にもとづき，すべての代替案に対して重視すべきその比重を数量化できる．従って，例えば Book3 が在庫切れなどであれば，次に重みの高い Book2 を選択することが適切であると判断できる．このような処理をエージェントに組み込むことで，エージェントは各状況に応じてより適切となり得る判断・行動を可能にすることが期待できる．

5. AHP 強化学習

5.1 AHP を用いるエージェントモデル

AHP 強化学習エージェントは，従来のエージェントモデルの学習器に，基礎知識である AHP 器を併用するように加えたものである．AHP 器を導入したエージェントのモデルを Fig.3 に示す．

AHP 器は，学習エージェント自身が設定目標を達

Table 5 各代替案の最終的な重み

	重み w
Book1	0.1924
Book2	0.3264
Book3	0.4812

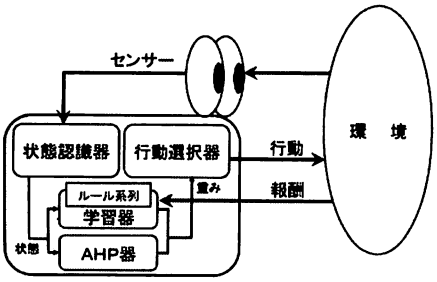

Fig.3 AHP 強化学習エージェントモデル

成するために本来備えておくべき基礎知識として設計され，状態認識器から与えられる情報に基づき，より適切な行動が優先されるように候補となる行動群を重み付けする．AHP により算出される重みと学習器の重みは，ある割合で合成され，行動選択器に送られる．エージェントは，合成された重みにもとづいてルーレット選択法により行動選択し，できるだけ適切な行動を出力することで環境との相互作用を通して学習する．

5.2 AHP 器の設計

上述したレスキュー問題における基礎知識を AHP で設計する際に，ＡＨＰ器の設計に必要となる事項である階層構造，代替案の重み付け，行動評価の増減の方針について記述する．

1. 階層構造

   レスキュー問題におけるエージェントは，被災者を探し抱えるタスクと，抱えた被災者を救急車に運び込むという２つのタスクを行う．被災者を探し抱えるためには，必ず被災者のところに行かなければならない．また，抱えた被災者を救急車に運びこむためには，必ず救急車のところに行かなければならない．したがって，エージェントにとって必要な基礎知識は，「被災者を抱えていないときに，被災者に近づく」知識と，「被災者を抱えているときに，救急車に近づく」知識である．これらの知識を AHP の階層構造で表すと，Fig.4 のようになる．

2. 代替案の重み付け

   上述した本の選定の例では，Table 2 に示す一対比較評価から人間の主観的判断に応じて値を選択し，一対比較表を作成したが，提案法ではその判断をシンプルに捉え，その作成を「段階的」な更新規定により自動化する．

   エージェントは，救急車，被災者，他のエージェント，障壁の位置情報を，環境の状態として認識する．エージェントが被災者や救急車に近づくためには，被災者や救急車の位置に関する情報が必要であるため，認識した被災者や救急車の位置情報を利用して，AHP 器の重みを更新する．

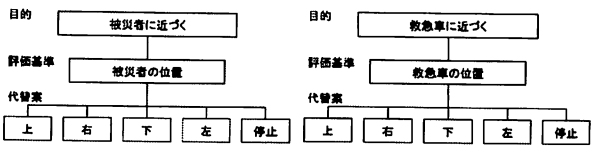

Fig.4 レスキュー問題に対する AHP の階層構造

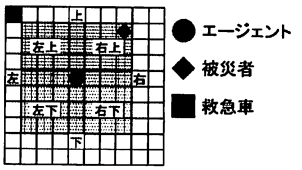

Fig.5 レスキュー問題におけるエージェントの視界

AHP 器は，エージェントを中心とする視野内において被災者，または救急車がどの場所に存在しているか（上・右上・右・右下・下・左下・左・左上）に応じ，代替案であるエージェントの行動（上・下・左・右・停止）を評価する．Fig.5 はエージェントの視野に関する図であり，エージェントから見て被災者が右上にいる場合の例である．

各エピソードの初期設定として，一対比較表のすべての値を Table 2 に示した「同程度重要」の 1 にセットする（Table 6）．その後の各ステップでは一対比較表は次のように更新される．エピソードのあるステップにおいて，例えば，エージェントから見た被災者の位置が右上である場合，現在の一対比較表で蓄えられている一対比較値を段階的に更新することにより代替案の右と上の重みの情報が増加するようにする．ここで「段階的」とは，Table 2 に示した整数値にしたがい，例えば一対比較値が 1 である場合，1 段階上げて 3 に更新することを指す．また，評価を下げる場合も同様に，ある一対比較値が 9 である場合は，1 段階下げて 7 に更新する．なお，一対比較表で利用される上限の値は Table 2 にしたがい 9 とし，下限値はその 9 の逆数 1/9 とする．このような数値の範囲を採ることで，一対比較表から算出される重みの秩序性が保たれる．Table 7 は，エージェントから見た被災者の位置が右上であった場合に，Table 6 から新たに更新される一対比較表である．

このように，AHP 器では，エージェントが観測する環境の各状態に基づいて，所定の規定の下で一対比較表の値が段階的に更新され，代替案の重みを算出する．

Table 6 一対比較表の初期設定

位置	上	下	左	右	停止	幾何平均 GM	重み w
上	1	1	1	1	1	1	0.2
下	1	1	1	1	1	1	0.2
左	1	1	1	1	1	1	0.2
右	1	1	1	1	1	1	0.2
停止	1	1	1	1	1	1	0.2

Table 7 被災者が右上にいる場合の一対比較表の例

位置	上	下	左	右	停止	幾何平均 GM	重み w
上	1	3	3	1	3	1.933	0.3333
下	1/3	1	1	1/3	1	0.644	0.1111
左	1/3	1	1	1/3	1	0.644	0.1111
右	1	3	3	1	3	1.933	0.3333
停止	1/3	1	1	1/3	1	0.644	0.1111

3. 行動評価の増減の方針

上述したように，一対比較表は各状態に応じて自動的に更新される．その自動的な更新に伴い，想定される状況の変化パターンに応じて，エージェントの各行動評価の増減の方針を決定する必要がある．その方針を次に示す．

・[エピソードの初期状態]
エピソードの初期では，「被災者に近づく」知識を使用する．視野内に被災者がいる場合は，被災者に近づく行動の評価を 1 段階上げ，近づかない行動の評価は 1 段階下げる．関係のない行動の評価は変更をしない．視野内に被災者がいない場合は，判断がつかないので各行動の評価は変更しない

・[被災者を抱えた場合]
AHP 器を Table 6 で示すように初期設定の状態にして，「救急車に近づく」知識を使用する．

・[被災者を救急車に運びこんだ場合]
AHP 器を Table 6 で示すように初期設定の状態にして，「被災者に近づく」知識を使用する．

・[エピソードの途中で被災者，あるいは救急車が見える状態から見えない状態になった場合]
今までに蓄積されてきた一対比較の判断がつきづらくなるため，高い評価（「同程度重要」よりも高い評価を指す）は 1 段階下げ，低い評価（「同程度重要」よりも低い評価を指す）は 1 段階上げることで，各行動の評価を「同程度重要」の一対比較値（1）に近づけるようにする．

AHP 器での代替案の重みは，基礎知識にもとづくエージェントの行動の重みの量として扱うことができ，エピソードの各ステップにおいてエージェントが観測する状態によって刻々と変化する．よって，AHP 器から得られる各行動の重みの役割は学習器の場合と似ている．このことから，エージェントは，たとえ学習器からの重みを利用しない場合であっても，AHP 器で得られる代替案の各行動の重みを用いてルーレット選択法により行動の選択も可能になる．

提案モデルでは，あくまでも，人間が与えた基礎知識にしたがい階層構造化された AHP 器により，エージェント自身が確率的な意思決定のもとで行動できる．よって，AHP 器による確率的な意思決定は，その行動自体が常に支配されるのでなく，人間が期待しなかった振舞いを実現する可能性が残されている．例えば，Table 7 を例にとると，被災者が右上にいる時でさえ，各行動の重みは，「右」「上」だけが与えられるのではなく，右と上以外の行動も重みとして与えられるため，ルーレット選択法により，右と上以外の行動が選択されることもある．このことから AHP 器は，さまざまな状況に応じてより適切となり得る行動を選択されやすくし，適切でない可能性が高い行動が選択されにくくなるように機能する．

5.3 AHP 器と学習器の重みの合成

AHP 強化学習エージェントは，2 章で示した Profit Sharing により学習を行う学習器と AHP 器の合成された重みを用いて行動選択を行っている．そのため，「基礎知識利用」か「学習による知識利用」かのジレンマが発生する．本論文では，AHP 器と学習器の各行動の重みを合成する方法として，「合成比減衰法」を用いる．合成比減衰法は，学習の初期段階では，良い性能とはいえない学習器の重みの利用を控え，AHP 器の重みを重視し，学習が進むにつれて，AHP 器の重みの利用を徐々に減衰させることにより，最終的には学習器のみの重みを利用する方法である．強化学習において，基礎知識に伴う行動の重みの利用が学習の長期に及ぶと，学習自体に悪影響を与えると共に，最終的に得られる学習の性能を阻害する．合成比減衰法は，知識の導入に伴う学習への悪影響を抑制する方法であることが報告されている[6]．

AHP 器と学習器の重みは下式のように合成する．

$$TWs = rate \cdot AHPWs + (1 - rate) \cdot LMWs$$

ここで，$rate(0 \leq rate \leq 1)$ は合成比，TWs は最終的に行動選択器に送られる各行動の重み，$AHPWs$ は AHP 器から得られる各行動の重み，$LMWs$ は学習器からの各行動の重みを表す．ただし，合成の計算を行う前に $AHPWs$ および $LMWs$ ともに行動群の合計がそれぞれ 1 になるようにする．

また，合成比減衰法は次式によって合成比 $rate$ を減衰させる．

$$rate = \alpha \cdot rate$$

ここで，$\alpha(0 \leq \alpha \leq 1)$ は減衰率であり，減衰は 1 エピソードごとに行う．なお，合成比 $rate$ と減衰率 α の

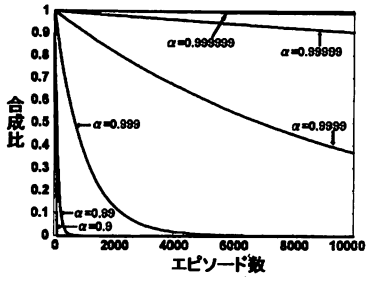

Fig.6 減衰率 α の違いによる合成比の減衰の比較

設定値によりさまざまな学習のバリエーションが可能となる．従来の強化学習エージェントは $rate = 0$ および $\alpha = 0$ とすることで実現できる．また，AHP 器からの重みのみを利用（つまり，全エピソードにおいて学習による知識を全く利用せず，基礎知識のみを利用）する場合は，$rate = 1$ および $\alpha = 1$ とすることで可能である．

なお，上述した減衰率 α の違いによる合成比減衰の傾向は Fig.6 のようになる．例えば減衰率 $\alpha = 0.999$ の曲線では，AHP 器の重みを利用する割合が約 6000 エピソード付近へ向けて徐々に低くなり，それに反して学習器の重みを利用する割合が徐々に増加する．それ以降は学習器の重みだけが利用され，従来の強化学習アルゴリズムと同等の処理になる．

6. 実験

3 章で記述したレスキュー問題を対象に，AHP 強化学習の視野の広さの違いによる学習性能を検討するために，以下に示す 3 つの実験を行う．

- 学習器のみを利用する従来の方法（主目標達成時に報酬を与える）MethodA と，AHP 器のみ（基礎知識のみ）を利用する方法で，視野の広さが学習器と AHP 器にどのような影響を与えるか観察する．
- 合成比減衰法による AHP 強化学習が視野の広さの違いから学習性能に受ける影響，減衰率の違いにより生じる学習性能の差を観察する．
- 複数タスク問題に対して副目標達成時に報酬を与えることで，従来法よりも高速な学習を実現した手法[12]MethodB に AHP を導入した方法で視野の広さの違いによる学習性能の検討を行う．

6.1 設定パラメータ

レスキュー問題の設定は，環境のサイズ $n = 15$，救急車の数 $N_{amb} = 1$，被災者の数 $N_v = 4$，エージェントの数 $N_A = 2$ とする．また，エージェントの視界 $m \times m$ の m を Depth と表す．Profit Sharing では，

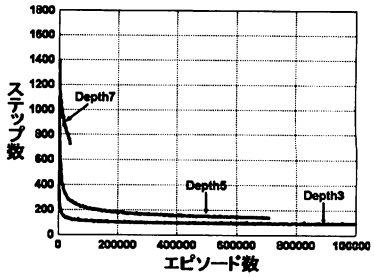

Fig.7 学習器のみを利用する方法の視野の広さの違いによる結果の比較

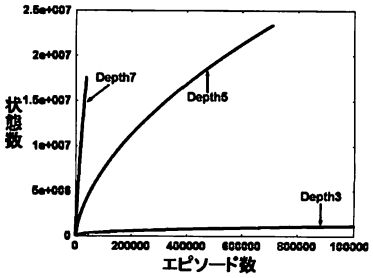

Fig.8 学習器のみを利用する方法の視野の広さの違いによる認識した状態数の比較

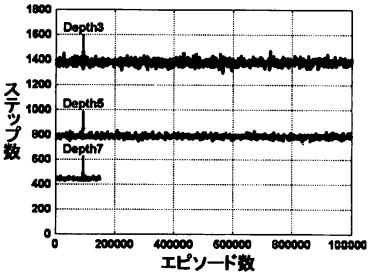

Fig.9 AHP器のみを利用する方法の視野の広さの違いによる結果の比較

初期のルール重みを 0.1，報酬割引率 $\beta = 0.9$，報酬 $r = 1.0$ とする．学習回数は 1000000 エピソードとする．ただし，以降に示すグラフ中で線が途中で切れているものは，その時点でメモリ不足となり終了したことを表す．MethodX+AHP($rate, \alpha$) は合成比 $rate$，減衰率 α の AHP を用いた MethodX を表すものとする．

6.2 視野の広さが学習器と AHP 器に与える影響

Fig.7 に従来法 MethodA，Fig.9 に AHP 器のみを利用する方法 AHP（1.0,1.0）の，視野の広さ Depth を 3,5,7 と変えたときの実験結果を示す．また，Fig.8 に MethodA と AHP(1.0,1.0) の，エージェントが認識した環境の状態の数をプロットした図を示す．Fig.7 より，学習器のみを利用する MethodA では，同じ設定のレスキュー問題に対して，視野の最も小さい Depth3 が Depth5 や 7 よりも良好な結果を示している．これは，Fig.8 に示すように，視野が大きいほどエージェントが認識する環境の状態の組合せの数（状態数）が多くなり，学習に時間を費やしているため，Depth5 や 7 は収束が遅くなり，学習性能の低下を招いたと考えら

れる．特に，Depth7 では，状態数が膨大になり，収束する前にメモリ不足となっている．

Fig.9 より，AHP 器のみを利用する AHP(1.0,1.0) では，視野が広いほど学習の性能が良くなっていることがわかる．これは，視野の広いほうが被災者や救急車の位置を速く特定しやすく，被災者や救急車が見えないという状態が減少したためと考えられる．

以上の実験結果より，AHP 器は，視野が広いほど，自分が置かれている環境の状態を把握しやすくなるため，学習の性能が上がることがわかる．しかし，学習器は，視野が広いほど，認識する状態の数が膨大になるため，学習に時間を費やし，性能が下がることがわかる．

6.3 AHP 強化学習の視野の広さの違いによる学習性能

Fig.10，11，12 に，MethodA に AHP 器を導入した AHP 強化学習 MethodA + AHP($rate, \alpha$) の Depth を 3,5,7 と変えたときの実験結果をそれぞれ示す．それぞれの図では，学習の初期と後期の傾向がわかるよう，x 軸のスケールを変えてプロットした 2 つの図を載せている．

Fig.10 より，学習の初期 500 エピソードあたりまでは MethodA + AHP(1.0,0.9) と MethodA + AHP(1.0,0.99) が MethodA よりも高速に学習を行っていることが観測できる．その後，MethodA + AHP(1.0,0.9) と MethodA + AHP(1.0,0.99) は，合成比の減衰によって，学習器のみ使用するようになることから，MethodA と同程度の学習性能を示している．減衰率が 0.999 以下の AHP 強化学習は基礎知識を多用したことで，学習性能が悪くなったと考えられる．よって，Depth3 では，基礎知識が有効に働く段階は，学習の非常に早い段階であるといえる．Depth3 では，視野が小さいために AHP の性能があまり発揮できないことと，状態数が少ないために比較的早く学習が行えることから，AHP を使用する期間は短いほうが良いといえる．

Fig.11 より，MethodA + AHP (1.0,0.99)，MethodA + AHP (1.0,0.999)，MethodA + AHP (1.0,0.9999) は，学習初期の合成比が大きいときには，MethodA よりも良い性能を示したが，基礎知識を利用する割合の減少とともに，MethodA と同程度または MethodA よりも悪い性能を示している．これは，基礎知識を利用することで学習の性能が上がっていたが，膨大な状態の数により，学習が完全に進行していないときに，基礎知識を利用する割合が小さくなったため，基礎知識を利用しない MethodA と同程度またはそれ以下になったと考えられる．

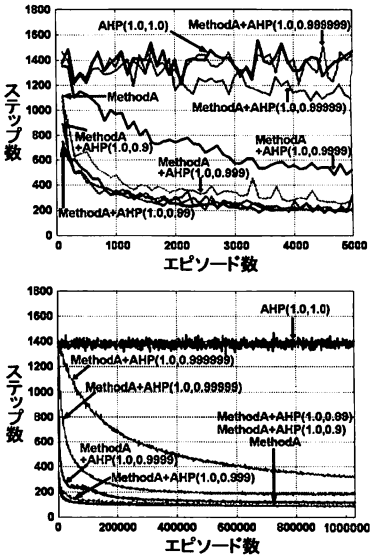

Fig.10 Depth3 のときの減衰率 α の違いによる AHP 強化学習の比較

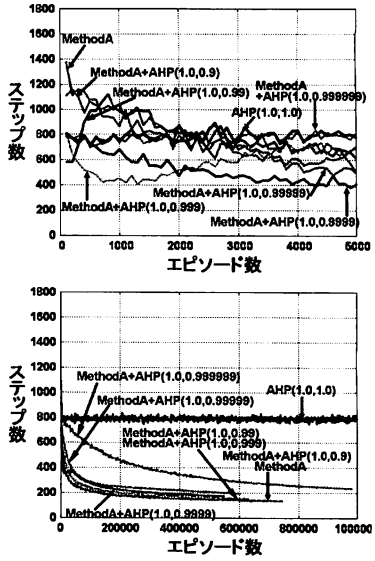

Fig.11 Depth5 のときの減衰率 α の違いによる AHP 強化学習の比較

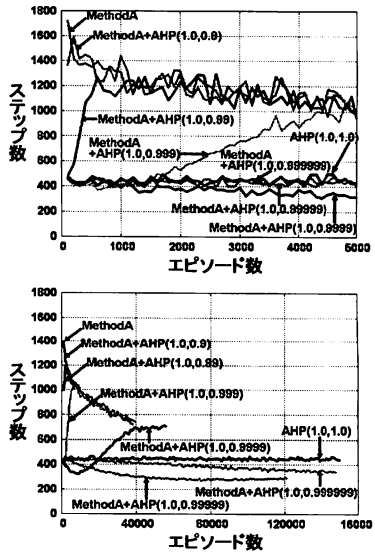

Fig.12 Depth7 のときの減衰率 α の違いによる AHP 強化学習の比較

Depth が 7 のときは状態数の多さがより顕著にわかる実験結果となっている．Fig.12 より，MethodA + AHP (1.0,0.9)，MethodA + AHP (1.0,0.99)，MethodA + AHP (1.0,0.999)，MethodA + AHP (1.0,0.9999) は，基礎知識を利用する割合が大きいときには，それぞれ MethodA と比べてはるかによい性能を示したにも関わらず，学習器のみを利用するようになった時点で，それぞれ MethodA と同程度まで学習の性能が落ちている．MethodA + AHP (1.0,0.9999) のグラフも基礎知識を使用する割合が小さくなると，性能が少し下がり，グラフが湾曲したように見える．

以上の実験結果より，AHP 強化学習では，Depth3 のように視野が狭い場合，AHP 器を導入することで，基礎知識を利用して，学習初期の収束を速めることが可能であるが，視野が広い場合，AHP 器自体の性能がよくても，膨大な状態数を学習する時間を必要とするため，学習に時間を費やすことがわかる．

6.4 MethodB に対する視野の広さの違いによる AHP 強化学習の学習性能

Fig.13, 14, 15 に，副目標達成時に報酬を与える方法である MethodB に対して，MethodA に対する実験と同様に，視野の広さ Depth を 3,5,7 と変えたときの実験結果をそれぞれ示す．MethodB の結果は，Depth3,5,7 すべての実験結果において，MethodA とほぼ同様の傾向を示している．また，複数タスク問題に対して副目標達成時に報酬を与えることで高速な学習を実現した MethodB は，本実験においても，同じ減衰率の MethodA と MethodB を比較した場合，MethodB の方が良い性能を示していることが観測できる．Fig.12, 15 の Depth7 の実験結果を比較すると，MethodA と MethodB の性能の差がよくわかる．また，MethodB は複数タスク問題に対して MethodA よりも高速な学習が可能であるため，学習初期において MethodB と AHP 強化学習の性能の差があまり大きくないものもある．例えば，Fig.10 の実験結果では，MethodA と MethodA + AHP (1.0,0.9) の初期の性能の差はよくわかるが，Fig.10 の実験結果では，MethodB と MethodB + AHP (1.0,0.9) の初期の性能の差はほとんどないことがわかる．

以上の実験結果より，Depth を 3,5,7 と変えたときの MethodB の傾向は，MethodA とほぼ同じような傾向を示すといえる．また，MethodB は MethodA よりも，高速な学習が可能であるが，視野が広い場合には，膨大な状態数の影響で，MethodA と同じく学習に時間を費やすといえる．

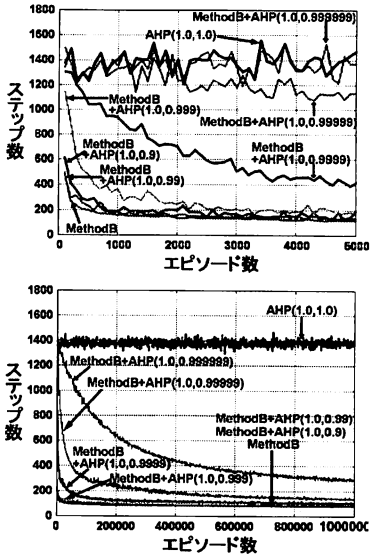

Fig.13 Depth3 のときの減衰率 α の違いによる AHP 強化学習（副目標達成時に報酬を与える）の比較

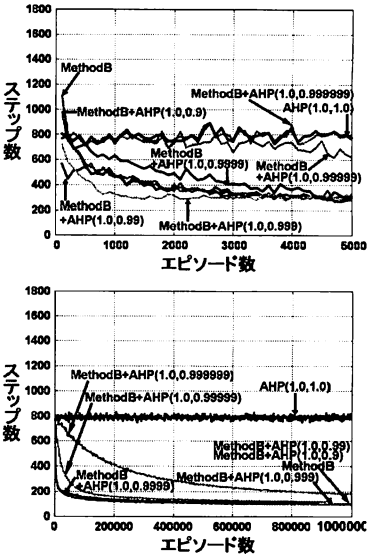

Fig.14 Depth5 のときの減衰率 α の違いによる AHP 強化学習（副目標達成時に報酬を与える）の比較

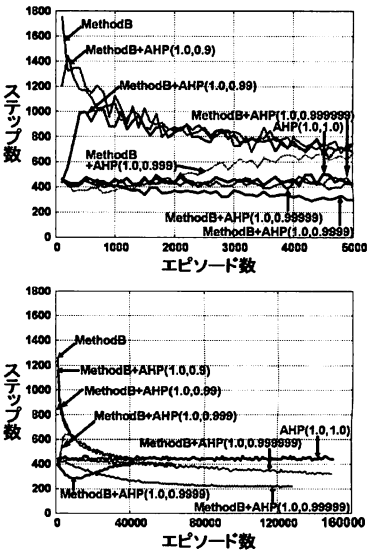

Fig.15 Depth7 のときの減衰率 α の違いによる AHP 強化学習（副目標達成時に報酬を与える）の比較

7. むすび

強化学習は，現実問題のような動的かつ不確実性を含む環境において，エージェントが有効に対応できる手法として期待されている．しかし，環境に対する情報を全く持たず，報酬だけを手がかりに学習を行うため，多くの学習時間を余儀なくされる場合が多く，現在のところ，実問題への応用や実際的な場面における利用において十分に対応できるとは言い難い．そのような問題に対処するため，我々は，学習エージェント自身が設定目標を達成するために本来備えておくべき基礎知識を階層化意思決定法（Analytic Hierarchy Process, AHP）で設計し，AHP 器として従来の強化学習エージェントへ導入する AHP 強化学習を提案している．本研究では，現実問題において多く存在する，複数タスクの問題を対象に，AHP 強化学習の視野の広さの違いによる学習性能の検討を行った．その結果，視野が狭い場合には，AHP 強化学習の性能を発揮できるが，視野が広い場合には，膨大な状態数の影響を受けて，学習に時間を費やすことを確認した．

参考文献

1) 荒井幸代, 宮崎和光, 小林重信, "マルチエージェント強化学習の方法論–Q-learning と Profit Sharing による接近–," 人工知能学会誌, Vol.13, No.5, pp.690–618, 1998.

2) 荒井幸代, "マルチエージェント強化学習–実用化に向けての課題・理論・諸技術との融合–," 人工知能学会誌, Vol.16, No.4, pp.476–481, 2001.

3) 荒井幸代, 田中信行, "マルチエージェント連続タスクにおける報酬設計の実験的考察–RoboCup Soccer Keepaway タスクを例として–," 人工知能学会誌, Vol.21, No.6, pp.537–546, 2006.

4) 伊藤昭, 金渕満, "知覚情報の粗視化によるマルチエージェント強化学習の高速化–ハンターゲームを例に–," 電子情報通信学会論文誌, (D-I), Vol.J84-D-I, No.3, pp.285–293, 2001.

5) Kaelbling, L. P., Littman, M. L., and Moore, A. W., "Reinforcement Learning : A Survey," Journal of Artificial Intelligence Research, Vol.4, pp.237–285, 1996.

6) 片山謙吾, 奥石尚宏, 成久洋之, "強化学習エージェントへの階層化意思決定方の導入–追跡問題を例に–," 人工知能学会論文誌, Vol.19, No.4, pp.279–291, 2004.

7) 加藤新吾, 松尾啓志, "動的環境下における Profit Sharing," 電子情報通信学会論文誌, (D-I), Vol.J84-D-I, No.7, pp.1067–1075, 2001.

8) 木村元, 宮崎和光, 小林重信, "強化学習システムの設計指針," 計測自動制御学会, 計測と制御, Vol.38, No.10, pp.618–623, 1999.

9) 宮崎和光, 木村元, 小林重信, "Profit Sharing に基づく強化学習の理論と応用," 人工知能学会論文誌, Vol.14, No.5, pp.800–807, 1999.

10) 宮崎和光, 荒井幸代, 小林重信, "Profit Sharing を用いたマルチエージェント強化学習における報酬分配の理論的考察," 人工知能学会誌, Vol.14, No.6, pp.1156–1164, 1999.

11) 西智樹, 高橋泰岳, 浅田稔, "モジュール型学習機構に置ける例示の理解に基づいた自律的なタスク分解," ロボティクス・メカトロニクス講演会 '05 予稿集, Vol.CD-ROM, 2P1-S-024, 2005.

12) 太田真由美, 金重徹, 片山謙吾, 南原英生, 成久洋之, "複数タスク問題に対するマルチエージェント強化学習の報酬発生タイミングと協調尺度," 第 19 回自律分散システム・シンポジウム資料, pp. 273–278, 2007.

13) 大内東, 山本雅人, 川村秀憲, "マルチエージェントシステムの基礎と応用," コロナ社, 2002.

14) Sutton, R. S. and Barto, A. G., "Reinforcement Learning : An Introduction," The MIT Press, Cambridge, MA, 1998. (邦訳 : 強化学習, 三上貞芳, 皆川雅章 共訳, 森北出版, 2000)

15) 田所諭, 北野宏明, 高橋友一, 松野文俊, 竹内郁雄, "RoboCup–Rescue 技術委員会 : RoboCup–Rescue 情報科学の緊急災害対応問題への挑戦," 情報処理学会誌, Vol.41, No.4, pp.412–418, 2000.

16) 高玉圭樹, "マルチエージェント学習–相互作用の謎に迫る–," コロナ社, 2003.

17) 内部英治, 浅田稔, 細田耕, "複数の学習するロボットの存在する環境における協調行動獲得のための状態空間の構成," 日本ロボット学会誌, Vol.20, No.3, pp.281–289, 2002.

18) 畝見達夫, "強化学習," 人工知能学会誌, Vol.9, No.6, pp.830–836, 1994.

19) 山村雅幸, 宮崎和光, 小林重信, "エージェントの学習," 人工知能学会論文誌, Vol.10, No.5, pp.683–689, 1995.

20) Weiss, G., "Multiagent Systems–Modern Approach to Distributed Artificial Intelligence–," The MIT Press, 1999.

Performance Comparison of AHP Reinforcement Learning by Difference of Depth of Recognition

Mayumi OHTA, Kengo KATAYAMA*, Hideo MINAMIHARA*
and Hiroyuki NARIHISA*

Graduate School of Engineering,
**Department of Information and Computer Engineering, Faculty of Engineering,*
Okayama University of Science
1-1 Ridai-cho, Okayama 700-0005, Japan

(Received October 1, 2007; accepted November 2, 2007)

Reinforcement Learning (RL) is a promising technique for creating agents that can be applied to real world problems. The most important features of RL are trial-and-error search and delayed reward. Thus, agents randomly act in the early learning state. However, such random actions are impractical for real world problems. Therefore, a design of practical reinforcement learning that can be learned in high speed has been desired. From this point of view, we have designed primary knowledge that humans intrinsically have in a process until a goal state is attained by using Analytic Hierarchy Process (AHP), and shown AHP Reinforcement Learning that integrates the primary knowledge as AHP module into standard RL algorithms.

The AHP module picks out necessary information from states of environment that agent recognized, and makes use of it in decision making such that agent has suitable actions. Therefore, the performance difference by depths that agent can recognize is expected. In this paper, we investigate the learning performance of the AHP-RL by the depth for the multi-task problem that exists much in real world.

Keywords: reinforcement learning; multi-agent; rescue problem; analytic hierarchy process.